**Q: Explain how the cartpole problem can be solved using the REINFORCE algorithm. Consider using pseudocode, UML, diagrams, or flowcharts to help illustrate your solution**.

The Cartpole problem can be solved using REINFORCE algorithm for this purpose we have to set a target value. The idea about reinforce is that we need to predict target values. The architecture is based on fully connected layers with the Relu activation function. The output layer is a fully connected layer with two outputs for each action.

n

Value 2

Value 1

state

target

The first plot on the left shows epsilon value each iteration. The right plot shows the epsilon function too.

**Q: Explain how the cartpole problem can be solved using the A2C algorithm. Consider using pseudocode, UML, diagrams, or flowcharts to help illustrate your solution.**

The Advantage Actor Critic (A2C) algorithm has two types of Reinforcement Learning algorithms (Policy Based and Value Based) together.

Chart, diagram, bubble chart

Description automatically generated

This diagram was taken from article, “Advantage Actor Critic Tutorial: minA2C”. Moreover, “The actor network outputs a probability distribution corresponding to each action. We sample actions from this probability distribution according to each action’s probability. If the action to go left has a value of .8 and the action to go right has a value of .2, we will only choose the left action 80% of the time and the right action 20% of the time. Because the output is a probability distribution, note that the agent will not always choose the action with the highest probability”. At the beginning of the learning process, the critic will likely make large errors causing the calculated TD error to be quite incorrect.

**Q: Explain how policy gradient approaches differ from value-based approaches, such as Q-learning**.

Both methods tends to be same, having same notations and concepts, but they are different internally. The difference is that how they approach action. Moreover, while using Q learning, the main purpose is to find max value by single action from set of actions. In the end, policy gradient can solve issues that value based can’t solve it. “A value-based method cannot solve an environment where the optimal policy is stochastic requiring specific probabilities, such as Scissor/Paper/Stone”.

**Q: Explain how actor-critic approaches differ from value- and policy-based approaches**.

The actor takes as input the state and outputs the best action. “It essentially controls how the agent behaves by learning the optimal policy (policy-based). The critic, on the other hand, evaluates the action by computing the value function (value based). Those two models participate in a game where they both get better in their own role as the time passes. The result is that the overall architecture will learn to play the game more efficiently than the two methods separately”.

Source:

<https://towardsdatascience.com/advantage-actor-critic-tutorial-mina2c-7a3249962fc8>

<https://theaisummer.com/Actor_critics/>